DOCUMENT RESUME

ED 194 600 TM 800 731

AUTHOR Olson, LeRoy A.
TITLE Writing Test Items. LES Papers on Learning and Teaching, Paper #50.
INSTITUTION Michigan State Univ., East Lansing. Learning and Evaluation Service.
PUB DATE [80]
NOTE 24p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Guidelines; Item Analysis; Multiple Choice Tests; *Objective Tests; *Test Construction; *Test Items; Writing Skills

ABSTRACT

A proficient item writer should possess several attributes: knowledge and understanding of the material being tested, continuous awareness of objectives, continous awareness of the instructional model, understanding of the students for whom the items are intended, skill in written communication, and skill in the techniques of item writing. The following general guidelines should be followed: (1) express items as clearly and simply as possible; (2) include all qualifications necessary to provide a reasonable basis for responding; (3) emphasize general tasks rather than small details; (4) avoid jargon and textbook language; (5) locate and delete irrelevant clues; (6) eliminate irrelevant sources of difficulty; (7) place all items of a given type together in a test; (8) prepare the key in advance of test administration; and (9) arrange for competent review of the items. Specific suggestions are presented for writing multiple-choice, alternative-response, or matching items. (Author/BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED194600

Writing Test Items

LeRoy A. Olson

Paper #50

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

L. Alexander

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

MSU IS AN AFFIRMATIVE ACTION/EQUAL OPPORTUNITY INSTITUTION

TM 800 731

TABLE OF CONTENTS

	PAGE
A. ESSENTIAL CHARACTERISTICS OF ITEM WRITERS	1
1. Knowledge and Understanding of the Material Being Tested	1
2. Continuous Awareness of Objectives	1
3. Continuous Awareness of Instructional Model	2
4. Understanding of the Students for Whom the Items are Intended	2
5. Skill in Written Communication	3
6. Skill in Techniques of Item Writing	3
B. GENERAL TIPS	3
1. Express Items as Precisely, Clearly and Simply as Possible	3
2. Include all Qualifications Necessary to Provide a Reasonable Basis for Responding	3
3. Emphasize General Tasks rather than Small Details	4
4. Avoid Jargon and Textbook Language	4
5. Locate and Delete Irrelevant Clues	4
6. Eliminate Irrelevant Sources of Difficulty	5
7. Place all Items of a Given Type Together in the Test	5
8. Prepare Keys or Model Answers in Advance of Test Administration	5
9. Arrange for Competent Review of the Items	5
C. WRITING SPECIFIC TYPES OF ITEMS	6
1. Multiple-Choice Items	6
a. State the Problem in the Stem	6
b. Include One Correct or Most Defensible Answer	7
c. Select Diagnostic Foils or Distractors	7
d. Options Should be Presented in a Logical, Systematic Order	8
e. Options Should be Grammatically Parallel and consistent with the Stem	8
f. Options Should be Mutually Exclusive	9
g. Insure that Correct Responses are not Consistently Shorter or Longer than the Foils	9
h. Eliminate Grammatical or Verbal Clues	10
i. Present the Problem in Novel Terms	10
j. Use Negatively-Stated Items Infrequently	11
k. Beware of "None of these," "None of the above," "All of these," and "All of the above"	11
l. Alter Item Difficulty by Making Options More Alike in Meaning	12
2. Alternative-Response Items	13
a. Include Only One Idea in Each Item	14
b. Eliminate Partly True-Partly False Items	14
c. Eliminate Specific Determiners	15
d. Insure that True and False Items are Approximately Equal in Length	15

PAGE

e. Balance the Number of True Items and False Items 15
f. Eliminate Vague Terms of Degree or Amount 16
g. Use Caution in Writing Negative-Stem Statements 16

3. Matching Items 17
a. Include Homogeneous Materials in Each Exercise 18
b. Include at Least Three to Five but no more than Eight to Ten Items in a Matching Set 18
c. Eliminate Irrelevant Clues 18
d. Place each Set of Matching Items on a Single Page 18
e. Reduce the Influence of Clues 19
f. Compose the Response List of Single Words or Very Short Phrases 19
g. Arrange the Responses in Systematic Order 19

D. THE PROOF OF THE ITEM WRITING IS IN THE ITEM ANALYSIS 20

WRITING TEST ITEMS

Writing test items is a matter of precision, perhaps more akin to computer programming than to writing prose. A test item must focus the attention of the examinee on the principle or construct upon which the item is based. Ideally, students who answer a test item incorrectly will do so because their mastery of the principle or construct in focus was inadequate or incomplete. Any characteristics of a test item which distract the examinee from the major point or focus of an item, reduce the effectiveness of that item. Any item answered correctly or incorrectly because of extraneous factors in the item, results in misleading feedback to both examinee and examiner.

A poet or writer, especially of fiction, relies on rich mental imagery on the part of the reader to produce an impact. For item writers, however, the task is to focus the attention of a group of students, often with widely varying background experiences, on a single idea. Such communication requires extreme care in choice of words and it may be necessary to try the items out before problems can be identified.

A. ESSENTIAL CHARACTERISTICS OF ITEM WRITERS

Given a task of precision communication, there are several attributes or mind sets that are characteristic of a proficient item writer.

1. Knowledge and Understanding of the Material Being Tested

At the University level, the depth and complexity of the material on which students are tested necessitates that only faculty members fully trained in a particular discipline can write concise, unambigious test items in that discipline. Further, the number of persons who can meaningfully critique test items, in terms of the principles or constructs involved, is limited. An agreement by colleagues to review each others' tests will likely improve the quality of items considerably prior to the first try-out with students.

2. Continuous Awareness of Objectives

A test must reflect the purposes of the instruction it is intended to assess. This quality of a test, referred to as content validity, is assured by specifying the nature and/or number of items prior to selecting and writing the items. Instructors sometimes develop a chart or test blueprint to help guide the selection of items. Such a chart may consider the modules or blocks of content as well as the nature of the skills a test is expected to assess.

In the case of criterion-referenced instruction, content validity is obtained by selecting a sample of criteria to be assessed. For content-oriented instruction, a balance may be achieved by selecting items in proportion to the amount of instructional time allotted to various blocks of material. An example of a test blueprint for a fifty-item test is shown below.

	Types of Tests	Reliability	Validity	Correlation	
Knowledge of terms	3	1	1	1	5
Comprehension of principles	3	4	4	4	15
Application of principles	2	4	6	5	17
Analysis of situations	1	2	2	2	7
Evaluation of solutions	-	2	2	2	6
	8	13	15	14	50

The blueprint specifies the number of items to be constructed for each cell of the two-way chart. For example, in the above test blueprint, four items are to involve the application of the principles of reliability.

3. Continuous Awareness of Instructional Model

Different instructional models require items of quite different characteristics for adequate assessment. For example, appropriate item difficulty in a mastery-model situation might be a maximum value of 20 (twenty-percent of the students answering incorrectly). On the other hand, items written for a normative model might have an average difficulty of the order of 30 to 40.

Ideally, item discrimination (the degree to which an item differentiates between students with high test scores and students with low test scores) should be minimal in a mastery-model situation. We would like to have all students obtain high scores. In the normative-model, item discrimination should be as high as possible in order that the total test differentiate among students to the maximum degree.

4. Understanding of the Students for Whom the Items are Intended

Item difficulty and discrimination are determined as much by the level of ability and range of ability of the examinees as they are by the characteristics of the items. Normative-model items must be written so that they provide the maximum intellectual challenge without posing a psychological barrier to student learning through excessive difficulty. In either the normative or mastery models, item difficulty must not be so low as to provide no challenge whatever to any examinee in a class.

It is generally easier to adjust the difficulty than to adjust the discrimination of an item. Item discrimination depends to a degree on the range of examinee

ability as well as on the difficulty of the item. It can be difficult to write mastery-model items which do not discriminate when the range of abilities among examinees is wide. Likewise, homogeneous abilities make it more difficult to write normative-model items with acceptably high discriminations.

No matter what the instructional model or the range of abilities in a class, the only way to identify appropriate items is to select items on the basis of subjective judgment, administer them, and analyze the results. Then only items of appropriate difficulty and discrimination may be retained for future use.

5. Skill in Written Communication

An item writer's goal is to be clear and concise. The level of reading difficulty of the items must be appropriate for the examinees. Wording must be more complicated than that used in instruction.

6. Skill in Techniques of Item Writing

There are many helpful hints and lists of pitfalls to avoid which may be helpful to the item writer. This is an area where measurement specialists may be particularly helpful. The remainder of this hand-out will be devoted to item-writing tips.

B. GENERAL TIPS

1. Express Items as Precisely, Clearly and Simply as Possible

Unnecessary material reduces the effectiveness of an item by forcing examinees to respond to the irrelevant material and perhaps be distracted by it. For example, the item

> In carrying out scientific research, the type of hypothesis which indicates the direction in which the experimenter expects the results to occur once the data has been analyzed is known as a(n)_____.

could be written

> An hypothesis which indicates the expected result of a study is called a(n)_____.

2. Include all Qualifications Necessary to Provide a Reasonable Basis for Responding

The item

> What is the most effective type of test item?

might be rewritten

> According to Ebel, the most versatile type of objective item for measuring a variety of educational outcomes is the_____.

The second version specifies whose opinion is to be used, narrows the task to consideration of objective items, and focuses on one item characteristic. The first version poses an almost impossible task.

3. Emphasize General Tasks Rather than Small Details

The item

> The product-moment coefficient of correlation was developed by
>
> 1. John Gosset
> 2. Sir Ronald Fisher
> 3. Karl Pearson

might be replaced by the item

> The product-moment coefficient of correlation is used to determine the degree of relationship between
>
> 1. two dichotomous variables.
> 2. a dichotomous variable and a continuous variable.
> 3. two continuous variables.

If an item on the product-moment coefficient of correlation is to be included in a test, it should concern some basic understanding or skill useful in determining when and how to apply the technique.

4. Avoid Jargon and Textbook Language

It is essential to use technical terms in any area of study. Sometimes, however, jargon and textbook phrases provide irrelevant clues to the answer, as in the following item.

> A test is valid when it
>
> 1. produces consistent scores over time.
> 2. correlates well with a parallel form.
> 3. measures what it purports to measure.
> 4. can be objectively scored.
> 5. has representative norms.

The phrase "measures what it purports to measure" is considered to be a measurement cliche which would be quickly recognized by students in the area. The item might be rewritten.

> The validity of a test may be determined by
>
> 1. measuring the consistency of its scores.
> 2. comparing its scores with those of a parallel form.
> 3. correlating its scores with a criterion measure.
> 4. inspecting the system of scoring.
> 5. evaluating the usefulness of its norms.

5. Locate and Delete Irrelevant Clues

Occasionally, verbal associations and grammatical clues render an item ineffective. For example, the item

> A test which may be scored merely by counting the correct responses is an _____ test.
>
> 1. consistent
> *2. objective
> 3. stable
> 4. standardized
> 5. valid

contains a grammatical incorsistency (an objective) which gives away the answer. The item could be rewritten

> A test which may be scored by counting the correct responses is said to be
>
> 1. consistent.
> *2. objective.
> 3. stable.
> 4. standardized.
> 5. valid.

6. Eliminate Irrelevant Sources of Difficulty

Other extraneous sources of difficulty may plague examinees in addition to the item faults mentioned above. Students may misunderstand the test directions if the test format is complex and/or the students are not familiar with it. When response keys are common to two or more items, care must be taken that students are made aware of the situation. If a set of items using a common key extends to a second page, the key should be repeated on the second page. Then students will not forget the key or have to turn back to an earlier page to consult the key.

Whenever complex or unfamiliar test formats are used, examinees should have an opportunity to practice responding to items prior to the actual test whose results are used for grading. Such a practice administration will also give the item writer an indication of difficulties students may be having with directions or with the test format.

7. Place all Items of a Given Type Together in the Test

Grouping like test items allows examinees to respond to all items requiring a common mind-set at one time. They don't have to continually shift back and forth from one type of task to another. Further, when items are grouped by type, each item is contiguous to its appropriate set of directions.

8. Prepare Keys or Model Answers in Advance of Test Administration

Preparing a key for objective type items or a model answer to essay or short answer items is an excellent way to check the quality of the items. If there are major flaws in items, they are likely to be discovered in the keying process. Preparing a model answer prior to administering the test is especially important for essay or other open-end items because it allows the examiner to develop a frame of reference prior to grading the first examination.

9. Arrange for Competent Review of the Items

Anyone who has attempted to proof his or her own copy knows that it is much better to have the material proofed by another person. The same principle applies to proofing test items. However, it is important that the outside reviewer be competent in the subject matter area. Unfortunately, critical review of test items is a demanding and time-consuming task. Item writers may make reciprocal

agreements with colleagues or may find advanced students to critique their items. Test construction specialists may provide helpful comments with respect to general item characteristics.

C. WRITING SPECIFIC TYPES OF ITEMS

The remainder of this handbook will deal with skills helpful in writing specific types of items. There is an almost infinite variety to the forms test items may take. Test items are often grouped into two main categories: objective items and constructed-response items. Objective items are those in which the examinee recognizes a best answer from options presented in the item. Objective items include multiple-choice items, alternative-response items and matching items. Constructed-response items include restricted-response items, short-answer items, completion items and essay items. Each type of item will be considered in turn on the following pages.

1. Multiple-Choice Items

A multiple-choice item presents a problem or question in the stem of the item and requires the examinee to select the best answer or option. The options consist of a most-correct answer and one or more distractors or foils. Consider the follwoing example.

> The statement "Attitude toward support of public schools is measured by performance at the polls" is an example of
>
> 1. a theory.
> 2. induction.
> 3. intuition.
> 4. an operational definition.
> 5. a deduction of "if then" statement.

The stem is the phrase "The statement 'Attitude toward support of public schools is measured by performance at the polls' is an example of." The numbered responses are the options, with option number four being the correct answer, and options one, two, three and five are foils or distractors. Now let us consider some hints for constructing this multiple-choice type of item.

a. State the Problem in the Stem

The item

> Multiple-choice items
>
> 1. may have several correct answers.
> 2. consists of a stem and some options.
> 3. always measure factual details.

does not have a problem or question posed in the stem. The examinee cannot determine the problem on which the item is focused without reading each of the items. The item should be revised, perhaps to read

> The components of a multiple-choice item are a
>
> 1. stem and several foils.
> 2. correct answer and several foils.
> 3. stem, a correct answer, and some foils.
> 4. stem and a correct answer.

A student who has been given the objective of recognizing the components of a multiple-choice item will read the stem, and immediately know the correct answer. The only remaining task is to locate the option which contains the complete list of components.

b. Include One Correct or Most Defensible Answer

The item below would be a good basis for discussion but probably should not be included in an examination.

> The most serious aspect of the energy crisis is the
>
> 1. possible lack of fuel for industry.
> 2. possibility of widespread unemployment.
> 3. threat to our environment from pollution.
> 4. possible increase in inflation.
> 5. cost of developing alternate sources of energy.

Such an item might be rewritten to focus on a more specific and/or agreed upon aspect of the energy crisis. It might also be written to focus on the opinion of a recognized expert:

> According to Professor Koenig, the most serious aspect of the energy crisis is the
>
> 1. possible lack of fuel for industry.
> 2. possibility of widespread unemployment.
> 3. threat to our environment from polution.
> 4. possible increase in inflation.
> 5. cost of developing alternative sources of energy.

c. Select Diagnostic Foils or Distractors Such as --

Cliches
Common Misinformation
Logical Misinterpretations
Partial Answers
Technical Terms or Textbook Jargon

The major purpose of a multiple-choice item is to identify examinees who do not have complete command of the concept or principle involved. In order to accomplish this purpose, the foils or distractors must appear as reasonable as the correct answer to students who have not mastered the material. Consider the

following item

A terminal may be defined as

1. a final stage in a computer program.
2. the place where a computer is kept.
3. an input-output device used when much interaction is required.
4. an auxilliary memory unit.
5. a slow but simple operating system.

Options 1 and 2 are derived from the common use of the word "terminal." They were each chosen by a number of students when the item was used in a pretest. Option 3 was keyed as the correct option.

d. Options Should be Presented in a Logical, Systematic Order

If a student who understands the principle being examined determines the correct answer after reading the item stem, then he or she should not have to spend time searching for that answer in a group of haphazardly arranged options. Option should always be arranged in some systematic manner. For example, dates of events should be arranged chronologically, numerical quantities in ascending order of size, and names in alphabetic order. Consider the following example.

What type of validity is determined by correlating scores on a test with scores on a criterion measured at a later date?

1. Concurrent
2. Construct
3. Content
4. Predictive

A student properly recognizing the description of predictive validity in the stem of the above item may go directly to the correct option since the options are in a logical order.

e. Options Should be Grammatically Parallel and Consistent with the Stem

Students are quick to take advantage of extraneous clues such as inconsistent stem and options. Thus they are responding to the item in terms of verbal skills possibly quite different from the skills the item is intended to measure. Note the extraneous clues in the item below.

A test which can be scored by a clerk untrained in the content area of the test is an

1. diagnostic test.
2. criterion-referenced tests.
3. objective test.
4. reliable test.
5. subjective test.

The examinee is led directly to option 3 by the last word in the stem, "an," which requires an option with its first word beginning with a vowel. Option 2 is rendered more implausible by the singular-plural inconsistency. The item might be rewritten as follows.

A test, which can be scored by a clerk untrained in the content area of the test, is said to be

1. diagnostic.
2. criterion-referenced.
3. objective.
4. reliable.
5. subjective.

f. Options Should be Mutually Exclusive

A knowledgeable examinee must be able to locate only one option which will contain the correct or best answer. Consider the faulty item below.

What should be the index of difficulty for an effective mastery-model test item?

1. Less than 10
2. Less than 20
3. More than 80
4. More than 90

If the index of difficulty is expressed as the proportion of the examinees who answer an item incorrectly, and option 1 is correct, then option 2 is also correct. The item should be rewritten as follows.

What should be the index of difficulty for an effective mastery-model test item?

1. Approximately 10
2. Approximately 20
3. Approximately 80
4. Approximately 90

g. Insure that Correct Responses are not Consistently Shorter or Longer than the Foils

If a test writer consistently writes correct options which are of different length than the foils or distractors, students will quickly learn to select correct answers on the basis of these idiosyncrasies. Longer correct options are perhaps most common since it is often necessary to add qualifiers to allow an option to be correct. For example:

A random sample is one in which

1. subjects are selected by levels.
2. each subject has an equal probability of being chosen for the sample.
3. every nth subject is chosen.
4. groups are the unit of analysis.

The item might be rewritten:

A random sample is one in which

1. subjects are selecged by levels in proportion to the number at each level in the population.
2. each subject has an equal probability of being chosen.
3. every nth subject is chosen from a list.
4. groups, rather than individuals, are the unit of analysis.

In the above revision, the correct option 2 is not conspicuously longer, as it was in the original version. In any case, shorter or longer correct options are not a problem unless they are consistently shorter or longer, so that students may establish a rule.

h. Eliminate Grammatical or Verbal Clues

Occasionally, a word is included in the stem of an item which furnishes the examinee with a verbal association with a word or words in the correct option.

The major purpose of item analysis is to

1. determine the distribution of test scores.
2. analyze the patterns of examinee responses.
3. determine whether the test content was appropriate.
4. find out if the test is reliable.
5. evaluate the overall difficulty of the test.

The word "analyze" in option 2 would be associated with "item analysis," leading examinees to the correct option through extraneous information. The problem could be solved, in this case, by changing the word "analyze" in option 2 to "consider." Further, an association might be provided for unknowledgeable examinees by changing option 3 to "analyze the content of the test."

i. Present the Problem in Novel Terms

Most often, we are interested in measuring complex mental processes such as application of principles rather than mere memory of factual knowledge. If we attempt to measure a complex mental task such as evaluation, but we use examples with which the students are familiar, we reduce the task to one of sheer memory.

In presenting examination items in novel terms, we must take care not to make the items extremely difficult relative to examples used in instruction. If we use novel item types, we must make sure that students understand the new process, and that they have had an opportunity to practice the required skill.

j. Use Negatively Stated Items Infrequently

There are situations in which the most meaningful task we can require of an examinee is to identify the exception in a set of options. The item stem will often ask:

> Which of the following is NOT an example of.....?

A major problem with a negatively-stated item is that students may miss the negation when reading the stem. A negatively-stated item does require an examinee to switch his or her mind set from that of looking for the best answer to that of locating the most definite non-answer. Items with negatively stated items can often be rewritten as effective positively-stated items.

For example, the negatively-stated item

> Which of the following is NOT a method of determining test reliability?
>
> 1. Coefficient of equivalence
> 2. Coefficient of stability
> 3. K-R #20
> 4. Split-halves procedure
> 5. Test-criterion intercorrelation

may be rephrased as a positively-stated item.

> Which of the following is a method of determining the validity of a test?
>
> 1. Coefficient of equivalence
> 2. Coefficient of stability
> 3. K-R #20
> 4. Split-halves procedure
> 5. Test-criterion correlation

The correct answer to each of the two above items is option 5.

k. Beware of "None of These," "None of the Above," "All of These," and "All of the Above."

The options "None of these" and "None of the above" should not be used when the examinee is to select the best, but not necessarily absolutely correct answer. They should only be used when one of the options would be agreed upon by experts as absolutely correct. The item below illustrates the inappropriate use of the "None of the above" option.

> What is an ideal level of difficulty for an objective test item?
>
> 1. 10
> 2. 20
> 3. 80
> 4. 90
> 5. None of the above

The correct answer would depend on the test model used as a frame of reference. The item might be improved by rewriting it as follows.

> What level of item difficulty would allow for maximum discrimination?
>
> 1. 10
> 2. 20
> 3. 50
> 4. 70
> 5. None of the above

Maximum discrimination could occur only at the level of difficulty of 50. Therefore, "None of the above" is a more appropriate option than in the earlier version of the item.

"All of these" and "All of the above" tend to be less useful options than the "None of these" type of option. When "All of these" or "All of the above" are used with a five-option multiple choice item, an examinee has only to recognize any two of the four options as correct to be led to the correct answer. Conversely, the examinee needs only to recognize one of the options as correct in order to reject the "All of the above" option.

If "None of the above" and "All of the above" are to be used as options, they must be used occasionally as the correct option. If they are seldom or never keyed as correct, the exmainees will soon recognize them as "fillers" and ignore them.

1. Alter Item Difficulty by Making Options More Alike or Less Alike in Meaning

Item options which are more alike in meaning provide a more difficult choice than do those which are more obviously different. Consider the three following items.

Easiest

> The quality of a test which indicates how consistently the test measures is called
>
> 1. objectivity.
> 2. reliability.
> 3. subjectivity.
> 4. validity.
>
> The correct option is number 2.

Harder

> The least expensive way to determine the reliability of a test is the
>
> 1. Kuder-Richardson procedure.
> 2. test-retest procedure.
> 3. parallel forms procedure.
> 4. parallel forms over time procedure.
>
> The correct option is number 1.

Hardest

Which of the following procedures provides the most stable estimate of equivalence?

1. K-R #20
2. K-R #21
3. Odd-even split halves
4. Randomized split-halves

The correct option is number 1.

2. ALTERNATIVE-RESPONSE ITEMS

An alternative-response item is a special case of the multiple-choice item format. There are many situations which call for either-or decisions, such as deciding whether a specific solution is right or wrong, whether to continue or to stop, whether to use singular or plural construction, and so on. For such situations, the alternative response item is an ideal measuring device. Since only two options are possible, alternative-response items are generally shorter, and, therefore, require less reading time. Students may respond to more alternative-response items than other types of items in a given length of time.

A major disadvantage of alternative-response items is the fact that students have fifty-fifty probability of answering the item correctly by chance alone. When carefully written and pretested, alternative-response items may be written which exceed multiple-choice items in ability to discriminate. Generally, however, it is considered necessary to have a larger number of alternative-choice items than of other types of items in order to achieve a given level of test reliability.

There are two main types of alternative-response items. One type is essentially a two option multiple-choice item. An example follows.

To determine the degree of relationship between two continuous variables, one must compute the

1. product-moment coefficient of correlation.
2. rank-order coefficient of correlation.

A second type of alternative-response item is the complete statement format, the most familiar example of which is the true-false item.

To determine the degree of relationship between two continuous variables, one must compute the rank-order coefficient of correlation.

1. True
2. False

The correct answer to the two-choice version is option 2, and the correct answer to the complete-statement version is option 2, False.

A major distinction between the complete-statement true-false type of item and items in multiple-choice or two-choice formats, is that the complete-statement item contains no criterion for answering the item. The criterion is outside of the item, rooted in the characteristics and experiences of each individual examinee. Each examinee must ask the question, true or false with respect to what? It follows that each complete-statement, true-false item must be unequivocally true or unequivocally false. It seems intuitive that proper wording and the elimination of extraneous clues are more crucial with the true-false type of item than with any other item format. Following are some points to consider in writing more effective alternative-response items.

a. Include Only One Idea in Each Item

Alternative-response items testing two ideas simultaneously provide little useful feedback to the examinee or to the examiner. If such "double-barrelled" items are answered incorrectly, it is not certain whether one of the ideas or both of them are responsible for the confusion. A two-idea item is given below.

Test validity is a function of test reliability, which can be improved by using fewer items.

1. True
2. False

The answer to the item is false since more items would be needed to improve test reliability. The item should be split into two true-false items.

b. Eliminate Partly True-Partly False Items

Lack of absolute truth or falsity of an item generally results from an item writer's failure to consider all possible frames of reference from which an item may be answered. Consider this example.

To be valid, a test must be content-balanced.

1. True
2. False

When considered from the point of view of an examiner of student achievement, the item is true. However, an examiner interested in student aptitude would likely consider the item false, since he or she would be primarily interested in accurate prediction of a criterion, not on the content of the test. The solution for the sample item above is simple.

To be valid, an achievement test must be content balanced.

1. True
2. False

The item may now be keyed true.

c. Eliminate Specific Determiners

In the real world, it is difficult to find things about which we can make absolute statements, such as "It is never true" or "It is always true." Very often, we need to qualify our statements, with words like often, as a rule, sometimes, may, and so on. Students use the following generalizations when answering true-false items. If an item contains an absolute specific determiner--such as always or never--mark the item false. If an item contains a qualifying specific determiner--such as sometimes or probably--mark it true.

Writers of alternative-response items must be extremely cautious in using words which may serve as specific determiners. In fact, item writers should attempt to use specific determiners in a way which will cause testwise but unknowledgeable students to answer the item incorrectly. An example of an inappropriate use of a specific determiner is given below.

All valid tests are reliable tests.

1. True
2. False

On the basis of the specific determiner, "all," an examinee would answer the item false, the keyed answer. A more appropriate version of the item is given below.

All valid aptitude tests are reliable tests.

1. True
2. False

In this case the specific determiner "all" would lead a testwise examinee to select the "false" answer, but the keyed answer is "true."

d. Insure that True and False Items are Approximately Equal in Length

Since we often have to qualify statements in order to make them unequivocally true, testwise students often use item length as an extraneous clue. Efforts should be made to write false items of about the same length as true items.

e. Balance the Number of True Items and False Items

The number of true and false items should be approximately balanced so that testwise students will approach an item on the basis of its content rather than on the probability of its being true or false. Some authors argue that the

proportion of false items should exceed the number of true items. They argue that, in general, respondents tend to agree with statements rather than disagree with them. Thus the false items will tend to be somewhat more difficult and the total test will be somewhat more reliable.

f. Eliminate Vague Terms of Degree or Amount

Words like "frequently" and "seldom" are especially open to interpretation in true-false items, which have no built-in frame of reference. It is generally possible to edit such vague terms out of true-false items. Here is an example.

> Reliability is frequently determined by the split-halves method.
>
> 1. True
> 2. False

This difficult-to-answer item may be rewritten as follows.

> The split-halves method is used to determine test reliability.
>
> 1. True
> 2. False

The revised item may be keyed 1. True.

g. Use Caution in Writing Negative Item Statements

The use of negative alternative-response item statements involves the same perils as does the use of negatives in multiple-choice item stems. Especially to be avoided is the double-negative, which sometimes pops up in alternative-response item statements. An example is given below.

> There is no advantage in not using specific determiners in alternative-response items.
>
> 1. True
> 2. False

The item should be rewritten.

> Specific determiners should be balanced between true and false items.
>
> 1. True
> 2. False

The item is now keyed 1. True.

3. MATCHING ITEMS

A matching item consists of two columns: one column of stems or problems to be answered, and another column of responses from which the answers are to be chosen. Traditionally, the column of stems is placed on the left and the column of responses is placed on the right. An example is given below.

> Directions: Match the data gathering procedures in the item column on the left with the name of the data gathered in the response column on the right. Place your answer in the blank to the left of each procedure. Each answer may be used only once.

Data Gathering Procedure	Type of Data
(a) 1. Administer two forms of a test	a. Coefficient of equivalence
(d) 2. Estimate reliability on the basis of item data	b. Coefficient of stability
(c) 3. Obtain odd-item and even-item scores	c. Internal consistency
(b) 4. Test, then retest in one month	d. Rational equivalence

Matching items are extensively used for matching terms with definitions, names with achievements, events with dates, and so on. A variation on the basic matching item is to use a diagram, chart, graph or map containing the stems or list of problems, and the names of the parts or components as the list of responses. An example of this variation is shown below.

		Item Components
(d) 1.	The components of a multiple-choice item are a	a. Correct answer(s)
		b. foil(s)
(b) 2.	1. stem and several foils.	c. option(s)
(b) 3.	2. correct answer and several foils.	d. stem(s)
(a) 4.	3. stem, a correct answer, and some foils.	
(b) 5.	4. stem and a correct answer	
(c) 6.		

Note that, in the above example, it is necessary to answer the multiple-choice item in order to answer the parent matching item. Note also that the responses (item components) in the list at the right have a (s) added to each response in order to eliminate singular-plural extraneous clues.

Because of the nature of the matching task, names with events, for example, it is clear that matching items often measure recognition of factual knowledge rather than higher level mental processes. Here are some hints for writing matching items.

a. Include Homogeneous Material in Each Exercise

Of all possible item types, matching items are the most likely to contain extraneous clues to the correct answers, especially when heterogeneous material is included in one matching item. Consider the following example.

(a) 1. Measures factual knowledge	a. Matching item
(b) 2. One stem and several options	b. Multiple choice item
(d) 3. Spread of scores around the mean	c. Reliability
(e) 4. Susceptible to specific determiners	d. Standard deviation
(c) 5. Test-retest score correlation	e. True-false item
(f) 6. Test score-criterion correlation	f. Validity

It is unlikely that even an examinee with only rudimentary knowledge would consider the options listing item types as responses to the stems on correlation. The above matching item is, in effect, operating as two very short matching items. The solution is to divide the item into two homogeneous items and add some stems and options to each set.

b. Include at Least Three to Five but no More than Eight to Ten Items in a Matching Set

Long sets of matching items require an examinee to do a good deal of clerical work in keeping track of stems and searching for options. Students with superior clerical skills will probably score higher than other students on long matching item sets. However, this is an extraneous skill to be avoided if possible. Another advantage of short matching items is that it is difficult to write long matching items which are homogeneous. Three to eight items per matching set is a reasonable compromise.

c. Eliminate Irrelevant Clues

Writing homogeneous matching item sets will go a long way toward eliminating irrelevant clues. Nevertheless, each stem-correct option pair should be carefully examined to make certain that there are no verbal association clues, plural-singular clues, and so on.

d. Place Each Set of Matching Items on a Single Page

No test item should be divided between pages when the test is mimeographed or printed. This is especially true for matching items where the student must do a considerable amount of searching for correct options. Having to flip pages while answering test items places an unnecessary, extraneous burden on the examinee.

e. Reduce the Influence of Clues and thereby Increase Matching Item Difficulty by

1. Using a different number of options than there are items, and
2. Allowing each option to be used more than once.

These rules are illustrated by the item revision shown below.

> Directions: Match the data gathering procedures in the item column on the left with the name of the data gathered in the response column on the right. Place your answer in the blank to the left of each procedure. Each response may be used more than once.

Data Gathering Procedure	Type of Data
(a) 1. Administer two forms of a test	a. Coefficient of equivalence
(d) 2. Estimate reliability on the basis of item data	b. Coefficient of stability
(c) 3. Obtain odd-item and even-item scores	c. Internal consistency
(b) 4. Test then retest in one month	d. Rational equivalence
	e. Standard deviation

f. Compose the Response List of Single Words or Very Short Phrases

Examinees will scan from left to right when reading a multiple-choice item, and they will invariably search for the correct response in the list on the right. Therefore, it is advisable to have the list on the right composed of single words or very short phrases. Compare the inconvenience of the negative example shown below with the convenience of the previous example in section e above.

Type of Data	Data Gathering Procedure
(a) 1. Coefficient of equivalence	a. Administer two forms of a test
(d) 2. Coefficient of stability	b. Estimate reliability on the basis of item data
(d) 3. Internal consistency	c. Obtain odd-item and even-item scores
(b) 4. Rational equivalence	d. Test, then retest in one month
	e. Test, then retest immediately

g. Arrange the Responses in Systematic Order: Alphabetical, Chronological, etc.

As was suggested for multiple-choice items, arrange the matching item responses in some logical order. Note that the responses in the item in the previous section are arranged alphabetically. This order enables examinees to find correct responses more quickly, thus reducing the amount of time they must spend on extraneous tasks and increasing the amount of time spent on the intended examination task.

D. THE PROOF OF THE ITEM WRITING IS IN THE ITEM ANALYSIS

Careful item writing can result in greatly improved tests. However, there can be no substitute for feedback to the item writer in the form of data from trial runs or actual usage. For even relatively small numbers of students, say twenty or thirty, useful data can be collected by means of optically-scanned sheets. Scan sheets and pencils may be picked up at the Scoring Office, 104 South Kedzie Hall.

After the test has been administered and the students have marked their responses to the items on the scan sheets, the sheets and the pencils are returned to the Scoring Office, along with a correct-answer key. The instructor must indicate to the Scoring Office supervisor that he or she wishes to receive an item analysis. An item analysis is generally available by noon of the next day after the sheets are delivered to the Scoring Office.

Mimeographed handouts are available from the Scoring Office which describe the item analysis print-out and how it may be used to evaluate items. Faculty may also call the Evaluation Unit of the Learning and Evaluation Service at 355-3408 for assistance from an Evaluation Consultant. Consultants are ready to assist faculty with planning, writing and evaluating their classroom tests, as well as tests for research purposes.